Statice

Anonymization and data privacy with Statice

Guide to GDPR compliance

Summary

01

Introduction

This guide presents the privacy protection mechanisms implemented in Statice's solution regarding the requirements of the General Data Protection Regulations (GDPR) for data anonymization.

In the first part, you can read about the legal requirements for data anonymization. We explain how the chosen technique should demonstrably protect against re-identification risks.

The second part presents how our protection mechanisms protect the data against privacy attacks and help you legally demonstrate the robustness of the anonymization process.

02

The anonymization process as per the GDPR

The GPDR defines anonymous data as «*information which do not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable.*»[1]

By extension, anonymization is the process of irreversibly transforming the data to reach that state. There is no single silver bullet approach to anonymization. Instead, anonymization processes aim to achieve a status where re-identification has become reasonably impossible.

The circumstances in which identification would no longer be possible vary. But it's possible to assess the likelihood of a re-identification happening in a given context based on parameters such as cost, time, or technology available to reverse a given anonymization technique.

The WP29 issued recommendations, highlighting three aspects crucial to evaluate the re-identification risks[2].

a) Are individuals' records in the dataset protected against singling out, where their record could be isolated from the data?

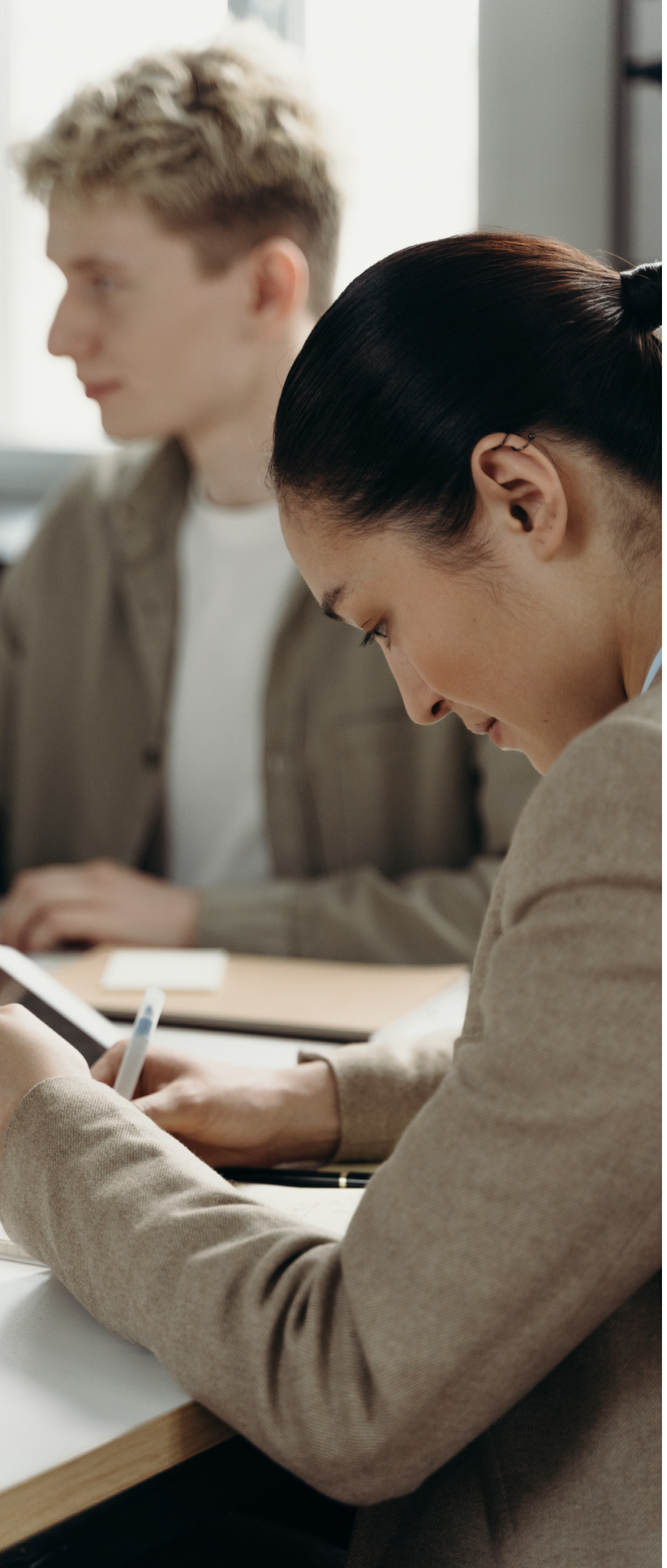
b) Are records in the dataset protected against cases of linkage, where (at least) two records that concern the same data subject could be linked in one or more datasets?

c) Are records in the dataset protected against cases of attribute inferences, where the value of a set of attributes might be deduced from the data.

According to WP29, an anonymization solution must offer protection against these three risks to be considered robust. Therefore, data protection and information security functions must assess and document the (un)likelihood of these scenarios to demonstrate the effectiveness of the anonymization process.

1 Recital 26 of the General Data Protection Regulation (GDPR)

2 Article 29 Working Party on the Protection of Individuals with regard to the Processing of Personal Data "Opinion 05/2014 on Anonymisation Techniques"



03

Comply with the requirements of the GDPR for anonymous data

Suppose the transformation process demonstrably protects against all three re-identification risks.

In that case, the anonymization will be considered robust against identifying a natural person by using all the means likely reasonable by the controller or a third party.

This robustness is the threshold required by the Article 2(a) of Directive 95/46/E[1] and the recital 26 of the GDPR to declare data anonymous.

Our solution offers two aspects to meet this definition.

First, we use data protection mechanisms more robust than the so-called «traditional» anonymization methods. Our anonymization method mitigates re-identification risks by breaking the one-to-one relationships with the original data.

Additionally, to mitigate re-identification risks we can make use of Differential Privacy, a strong privacy-enhancing technique, in the synthetization process.

Secondly, we provide privacy evaluators that let you measure and document the singling out, linkability, and inference risks.

With these assessments, you can quantify the residual risk and demonstrate that the re-identification is “reasonably” impossible, ensuring that your synthetic data is safe and legally compliant with the GDPR requirements for data anonymization.

1 Article 2(a) of Directive 95/46/E of the European Parliament and of the council on the protection of individuals with regard to the processing of personal data and on the free movement of such data

04

Lower the re-identification risks with synthetic data

Statice's anonymization solution relies on the process of synthetic data generation. Our technology can generate algorithmically synthetic data that looks and behaves like real data.

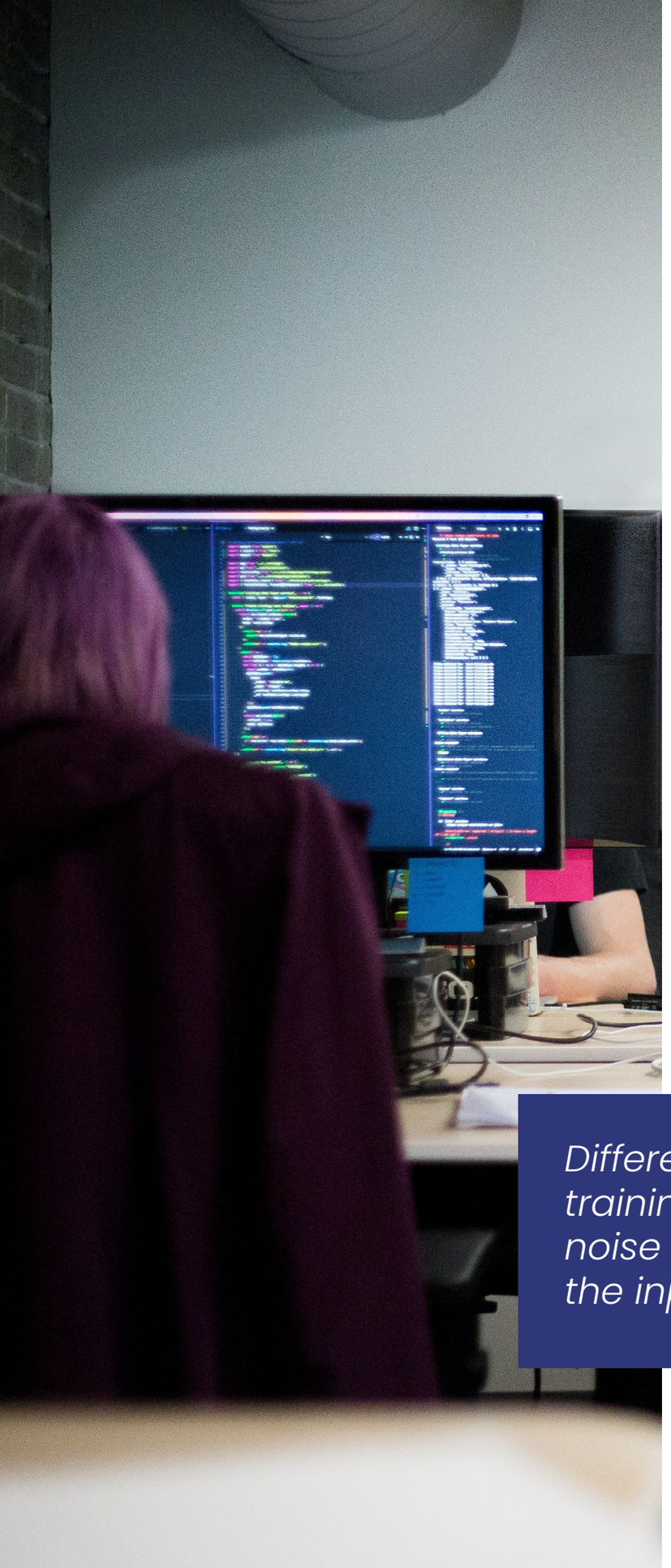
Under the hood, models learn the statistical distribution of the original dataset and draw artificial samples from it to generate a synthetic dataset.

This process demonstrably breaks the one-to-one relationship between original and synthetic data, minimizing the risk of re-identification.

Contrary to other anonymization methods, synthetic data generation is irreversible: there is no key to returning from the synthetic records to the original ones.

Additionally, this method is customizable to adapt to the data disclosure context. For example, in a context where data would be made available to a third party, it would be primordial to ensure that the highest level of privacy is maintained.

The synthetic data generation process demonstrably breaks the one-to-one relationship between original and synthetic data, minimizing the risk of re-identification.



05

Protect model privacy with Differential Privacy

Synthetic data generation mitigates re-identification risks. Additionally, it's possible to use Differential Privacy to ensure that the machine learning models used to generate the synthetic data do not breach privacy.

Differential Privacy (DP) is a mathematical definition of privacy widely accepted by the community.

To this day, differentially private systems have never been successfully attacked. They protect features in the training data from being memorized by the model, using noise to mask the presence of any particular individual in the input data.

Differentially private systems protect features in the training data from being memorized by the model using noise to mask the presence of any particular individual in the input data.

06

Assess re-identification risks with Statice privacy evaluators

Once you have generated synthetic data, we provide metrics and empirical privacy evaluations to measure the re-identification risk.

Unlike with other anonymization techniques like k-anonymity, there is no direct link between synthetic and original records. However, it would be wrong to assume that synthetic data is exempt from any linkability risk.

The Linkability Evaluator measures this risk by evaluating how much help the synthetic data gives to an attacker who wants to establish links between records belonging to the same individual.

The Inference Evaluator detects specific information leaks from the synthetic records, which can only be learned from the original data. Specifically, it analyzes the robustness of the synthetic data against inference attacks in which a knowledgeable adversary possesses information on a subset of the original records and attempts to reconstruct the unknown attributes by linking their partial knowledge with the synthetic data.

The Singling out Evaluator analyzes the probability of isolating records that identify an individual. It measures the robustness of the dataset against an attack in which an attacker would use the synthetic data to determine the presence of an individual with one or more given attributes in the dataset.

Combined, these evaluations provide a data-driven assessment of the protection against re-identification risks. Satisfactory results guarantee that you have a robust anonymization process. It also demonstrates that the re-identification risk threshold is minimal, ensuring that your data is safe and legally compliant with the GDPR requirements.

“Statice anonymization is based on synthetic data and differential privacy. No original data is reproduced in the synthetic data, only the statistical properties of the original data.

Differential privacy provides further protection through noise injection. Following the idea of factual anonymization of GDPR, it can be constituted that Statice anonymization provides extremely strong protection against re-identification.”

Marc Leutsch
Data Protection Officer

Discover our Re-identification Risk Assessment Framework for Data Protection Impact Assessment
Get in touch with us!

About

Statice

Statice develops state-of-the-art data privacy technology that helps companies double-down on data-driven innovation while safeguarding the privacy of individuals.

With Statice, companies generate privacy-preserving synthetic data compliant for any type of data integration, processing, and dissemination. Enterprises from the financial, insurance, and healthcare industries drive data agility and unlock the creation of value along their data lifecycle.

Start training your machine learning models or share data internally or with partners thanks to Statice.

- Recital 26 of the General Data Protection Regulation (GDPR)
- Article 29 Working Party on the Protection of Individuals with regard to the Processing of Personal Data "Opinion 05/2014 on Anonymisation Techniques"
- Article 2(a) of Directive 95/46/E of the European Parliament and of the council on the protection of individuals with regard to the processing of personal data and on the free movement of such data
- European Data Protection Supervisor on Anonymization
- European Medicine Agency, External guidance on the implementation of the European Medicines Agency policy on the publication of clinical data for medicinal products for human use.

Contact us

www.statice.ai
hello@statice.ai

Find us